데이터 과학 용어 요약집

1. A/B 테스트(A/B Testing): 두 가지 버전의 제품, 웹페이지 또는 모델을 비교하여 어떤 버전이 더 나은 성능을 보이는지 결정하는 데 사용되는 통계적 방법입니다.
2. 정확도(Accuracy): 분류 모델이 평가하는 모든 인스턴스 중에서 결과를 얼마나 자주 올바르게 예측하는지 측정하는 지표입니다.
3. 아다부스트(Adaboost): 약한 분류기를 결합하여 강력한 분류기를 만드는 앙상블 학습 알고리즘입니다.
4. 알고리즘(Algorithm): 컴퓨터가 문제를 해결하거나 작업을 수행하기 위해 따르는 단계별 지침 또는 규칙의 집합입니다.
5. 분석(Analytics): 의미 있는 통찰력을 추출하기 위해 데이터를 해석하고 검토하는 과정입니다.
6. 이상 감지(Anomaly Detection): 데이터에서 비정상적인 패턴이나 특이값을 식별하는 것입니다.
7. ANOVA(분산 분석 - Analysis of Variance): 샘플에서 그룹 평균 간의 차이를 분석하는 데 사용되는 통계적 방법입니다.
8. API(응용 프로그래밍 인터페이스 - Application Programming Interface): 하나의 소프트웨어 애플리케이션이 다른 애플리케이션과 상호 작용할 수 있도록 하는 일련의 규칙입니다.
9. AUC-ROC(ROC 곡선 아래 면적 - Area Under the ROC Curve): 분류 모델이 양성 또는 음성 예측으로 간주하는 방법을 고려하여 전체적으로 얼마나 잘 수행되는지 알려주는 지표입니다.
10. 배치 경사 하강법(Batch Gradient Descent): 전체 훈련 데이터셋을 사용하여 모델 파라미터를 업데이트하는 최적화 알고리즘입니다(미니 배치 경사 하강법과 다름).
11. 베이즈 통계(Bayesian Statistics): 사전 지식과 관찰된 데이터를 결합하는 통계적 접근 방식입니다.
12. BI(비즈니스 인텔리전스 - Business Intelligence): 조직이 정보에 기반한 비즈니스 의사결정을 내리는 데 도움이 되는 기술, 프로세스 및 도구입니다.
13. 편향(Bias): 모델에서 발생하는 오류로, 실제 값에서 벗어난 값을 일관되게 예측하게 합니다.
14. 편향-분산 트레이드오프(Bias-Variance Trade-off): 모델에서 편향과 분산으로 인해 발생하는 오류 간의 균형입니다.
15. 빅 데이터(Big Data): 기존의 데이터 처리 방법으로는 쉽게 처리할 수 없는 크고 복잡한 데이터셋입니다.
16. 이진 분류(Binary Classification): 데이터를 스팸 또는 스팸 아님과 같이 두 그룹으로 분류하는 것입니다.
17. 부트스트랩 샘플링(Bootstrap Sampling): 데이터셋에서 복원 추출 방식으로 무작위 샘플을 추출하는 재샘플링 기법입니다.
18. 범주형 데이터(Categorical data): 범주 또는 그룹을 나타내며 제한적이고 고정된 수의 고유 값을 가질 수 있는 변수입니다.
19. 카이제곱 검정(Chi-Square Test): 두 범주형 변수 사이에 유의미한 연관성이 있는지 확인하는 데 사용되는 통계 테스트입니다.
20. 분류(Classification): 데이터 포인트를 미리 정의된 클래스 또는 그룹으로 분류하는 것입니다.
21. 클러스터링(Clustering): 특정 기준에 따라 유사한 데이터 포인트를 함께 그룹화하는 것입니다.
22. 신뢰 구간(Confidence Interval): 특정 신뢰 수준으로 모수의 참값을 추정하는 데 사용되는 값의 범위입니다.
23. 혼동 행렬(Confusion Matrix): 분류 알고리즘의 성능을 평가하는 데 사용되는 표입니다.
24. 상관 관계(Correlation): 두 변수 간의 연관성 정도를 나타내는 통계적 측정값입니다.
25. 공분산(Covariance): 두 확률 변수가 함께 얼마나 변하는지를 측정하는 값입니다.
26. 교차 엔트로피 손실(Cross-Entropy Loss): 분류 문제에서 일반적으로 사용되는 손실 함수입니다.
27. 교차 검증(Cross-Validation): 데이터를 훈련 및 테스트를 위한 여러 하위 집합으로 분할하여 모델의 성능을 평가하는 기술입니다.
28. 데이터 클리닝(Data Cleaning): 데이터셋에서 오류 또는 불일치를 식별하고 수정하는 과정입니다.
29. 데이터 마이닝(Data Mining): 대규모 데이터셋에서 가치 있는 패턴 또는 정보를 추출하는 것입니다.
30. 데이터 전처리(Data Preprocessing): 원시 데이터를 분석에 적합한 형식으로 정리하고 변환하는 것입니다.
31. 데이터 시각화(Data Visualization): 이해를 돕기 위해 데이터를 그래프 또는 시각적 형식으로 표현하는 것입니다.
32. 결정 경계(Decision Boundary): 분류 문제에서 다른 클래스를 분리하는 구분선입니다.
33. 결정 트리(Decision Tree): 일련의 규칙을 기반으로 의사결정을 내리는 트리 모양의 모델입니다.
34. 차원 축소(Dimensionality Reduction): 중요한 정보를 유지하면서 데이터셋의 특징(feature) 수를 줄이는 것입니다.
35. 고유값 및 고유벡터(Eigenvalue and Eigenvector): 선형 대수에서 사용되는 개념으로, 복잡한 데이터셋을 변환하고 단순화하기 위해 차원 축소에 자주 사용됩니다.
36. 엘라스틱 넷(Elastic Net): L1 및 L2 페널티를 결합한 정규화 기법입니다.
37. 앙상블 학습(Ensemble Learning): 여러 모델을 결합하여 전체 성능과 정확도를 향상시키는 것입니다.
38. 탐색적 데이터 분석(EDA - Exploratory Data Analysis): 데이터의 특성과 관계를 이해하기 위해 데이터를 분석하고 시각화하는 것입니다.
39. F1 점수(F1 Score): 분류 모델에서 정밀도(precision)와 재현율(recall)을 결합한 지표입니다.
40. 오류 양성 및 오류 음성(False Positive and False Negative): 이진 분류에서의 잘못된 예측입니다.
41. 특징(Feature): 예측을 위해 ML 모델의 입력으로 사용되는 데이터 열입니다.
42. 특징 공학(Feature Engineering): 기존 특징에서 새로운 특징을 생성하여 모델 성능을 향상시키는 것입니다.
43. 특징 추출(Feature Extraction): 중요한 특징을 선택하여 데이터의 차원을 줄이는 것입니다.
44. 특징 중요도(Feature Importance): 모델 예측에 대한 각 특징의 기여도를 평가하는 것입니다.
45. 특징 선택(Feature Selection): 모델에 가장 관련성이 높은 특징을 선택하는 것입니다.
46. 가우시안 분포(Gaussian Distribution): 통계 모델링에서 자주 사용되는 확률 분포 유형입니다.
47. 지리 공간 분석(Geospatial Analysis): 지리적 데이터 내의 패턴과 관계를 분석하고 해석하는 것입니다.
48. 경사 부스팅(Gradient Boosting): 약한 모델이 순차적으로 훈련되어 이전 모델의 오류를 각기 수정하는 앙상블 학습 기법입니다.
49. 경사 하강법(Gradient Descent): 모델의 파라미터를 조정하여 오류를 최소화하는 데 사용되는 최적화 알고리즘입니다.
50. 그리드 서치(Grid Search): 가능한 모든 조합에서 모델을 평가하여 하이퍼파라미터를 튜닝하는 방법입니다.
51. 이분산성(Heteroscedasticity): 회귀 모델에서 오류의 가변성이 일정하지 않은 것입니다.
52. 계층적 클러스터링(Hierarchical Clustering): 클러스터 분석 방법 중 하나로, 데이터를 트리와 같은 클러스터 구조로 구성하여 트리의 각 수준에서 다양한 데이터 포인트 그룹 간의 관계 및 유사성을 보여줍니다.
53. 하이퍼파라미터(Hyperparameter): 훈련 프로세스가 시작되기 전에 값이 설정되는 파라미터입니다.
54. 가설 검정(Hypothesis Testing): 샘플 데이터를 기반으로 모집단 모수에 대한 가설을 테스트하는 통계적 방법입니다.
55. 결측치 대체(Imputation): 다양한 기술을 사용하여 데이터셋의 누락된 값을 채우는 것입니다.
56. 추론 통계(Inferential Statistics): 데이터 샘플을 기반으로 모집단에 대한 추론을 포함하는 통계학의 한 분야입니다.
57. 정보 이득(Information Gain): 의사결정 트리에서 데이터 분류에 대한 특징의 효율성을 평가하는 데 사용되는 측정값입니다.
58. 사분위 범위(IQR - Interquartile Range): 통계적 산포도 측정값으로, 첫 번째 사분위수와 세 번째 사분위수 사이의 범위를 나타냅니다.
59. 조인트 플롯(Joint Plot): Seaborn에서 두 변수 간의 관계와 개별 분포를 탐색하는 데 사용되는 데이터 시각화 유형입니다.
60. 결합 확률(Joint Probability): 두 개 이상의 사건이 동시에 발생할 확률로, 통계 분석에서 자주 사용됩니다.
61. 주피터 노트북(Jupyter Notebook): 라이브 코드, 방정식, 시각화 및 서술 텍스트가 포함된 문서를 생성하고 공유하기 위한 오픈 소스 웹 애플리케이션입니다.

62. K-평균 클러스터링(K-Means Clustering): 데이터셋을 구별되고 겹치지 않는 하위 집합으로 분할하기 위한 인기 있는 알고리즘입니다.
63. K-최근접 이웃(KNN - K-Nearest Neighbors): 새로운 데이터 포인트가 다른 데이터 포인트와 얼마나 가까운지에 기반한 간단하고 널리 사용되는 분류 알고리즘입니다.
64. L1 정규화(L1 Regularization): 손실 함수에 계수의 절댓값을 페널티 항으로 추가하는 것입니다.
65. L2 정규화(릿지 - L2 Regularization(Ridge)): 손실 함수에 계수의 제곱값을 페널티 항으로 추가하는 것입니다.
66. 선형 회귀(Linear Regression): 종속 변수와 하나 이상의 독립 변수 간의 관계를 모델링하는 통계적 방법입니다.
67. 로그 우도(Log Likelihood): 우도 함수의 로그로, 최대 우도 추정에서 자주 사용됩니다.
68. 로지스틱 함수(Logistic Function): 이진 결과의 확률을 모델링하는 데 로지스틱 회귀에서 사용되는 시그모이드 함수입니다.
69. 로지스틱 회귀(Logistic Regression): 이진 결과의 확률을 예측하기 위한 통계적 방법입니다.
70. 머신러닝(Machine Learning): 시스템이 데이터로부터 학습하고 예측할 수 있도록 하는 인공지능의 하위 분야입니다.
71. 평균 절대 오차(MAE - Mean Absolute Error): 예측값과 실제 값 간의 평균 절대 차이를 측정하는 지표입니다.
72. 평균 제곱 오차(MSE - Mean Squared Error): 예측값과 실제 값 간의 평균 제곱 차이를 측정하는 지표입니다.
73. 평균(Mean): 일련의 숫자의 평균값입니다.
74. 중앙값(Median): 정렬된 숫자 집합의 중간 값입니다.
75. 지표(Metrics): 정확도, 정밀도, 재현율, F1 점수 등 머신러닝 모델의 성능을 평가하는 데 사용되는 기준입니다.
76. 모델 평가(Model Evaluation): 다양한 지표를 사용하여 머신러닝 모델의 성능을 평가하는 것입니다.
77. 다중 공선성(Multicollinearity): 회귀 모델에서 독립 변수 간의 높은 상관 관계가 존재하는 것입니다.
78. 다중 레이블 분류(Multi-Label Classification): 단일 입력에 여러 레이블을 할당하는 것으로, 단 하나의 레이블만 할당하는 것과 대조됩니다.
79. 다변량 분석(Multivariate Analysis): 여러 변수를 사용하여 변수 간의 관계를 이해하는 데이터 분석입니다.
80. 나이브 베이즈(Naive Bayes): 분류에 사용되는 베이즈 정리를 기반으로 하는 확률적 알고리즘입니다.
81. 정규화(Normalization): 숫자 변수를 표준 범위로 스케일링하는 것입니다.
82. 귀무 가설(Null Hypothesis): 관찰된 결과와 예상된 결과 사이에 유의미한 차이가 없다고 가정하는 통계적 가설입니다.
83. 원-핫 인코딩(One-Hot Encoding): 머신러닝 모델을 위해 범주형 변수를 이진 행렬로 변환하는 기술입니다.
84. 순서형 변수(Ordinal Variable): 의미 있는 순서가 있지만 반드시 등간격을 가지지 않는 범주형 변수입니다.
85. 특이값(Outlier): 데이터셋의 다른 관측치와 현저하게 벗어나는 관측치입니다.
86. 과적합(Overfitting): 훈련 데이터에서는 잘 작동하지만 새롭고 보지 못한 데이터에서는 성능이 저조한 모델입니다.
87. 판다스(Pandas): 구조화된 데이터를 작업하기 위한 파이썬의 표준 데이터 조작 라이브러리입니다.
88. 피어슨 상관 계수(Pearson Correlation Coefficient): 두 변수 간의 선형 관계를 측정하는 지표입니다.
89. 푸아송 분포(Poisson Distribution): 고정된 시간 또는 공간 간격 내에서 주어진 수의 사건이 발생할 확률을 나타내는 이산 확률 분포입니다.
90. 정밀도(Precision): 분류 모델이 만든 총 양성 예측 수에 대한 실제 양성 예측 수의 비율입니다.
91. 예측 분석(Predictive Analytics): 데이터, 통계 알고리즘 및 머신러닝 기술을 사용하여 미래 결과의 가능성을 식별하는 것입니다.
92. 주성분 분석(PCA - Principal Component Analysis): 데이터를 새로운 특징 프레임워크로 변환하여 정보를 단순화하면서 기본 패턴을 보존하는 차원 축소 기술입니다.
93. 주성분(Principal Component): 주성분 분석에서 데이터셋에서 가장 많은 분산을 포착하는 축입니다.
94. P-값(P-value): 가설 검정 중 관찰된 결과만큼 극단적이거나 그보다 더 극단적인 결과를 얻을 확률입니다.
95. Q-Q 플롯(Quantile-Quantile Plot): 데이터셋이 특정 이론적 분포를 따르는지 평가하기 위한 그래픽 도구입니다.
96. 분위수(Quantile): 데이터셋을 동일한 부분으로 나누는 데이터 포인트 또는 데이터 포인트 집합입니다.
97. 랜덤 포레스트(Random Forest): 다수의 결정 트리를 구성하고 이를 병합하여 더 정확하고 안정적인 예측을 제공하는 앙상블 학습 방법입니다.
98. 무작위 샘플(Random Sample): 모집단의 각 구성원이 선택될 동일한 기회를 갖는 샘플입니다.
99. 확률 변수(Random Variable): 가능한 값이 무작위 현상의 결과인 변수입니다.
100. 재현율(Recall): 분류 모델에서 실제 양성 인스턴스 총수에 대한 실제 양성 예측 수의 비율입니다.
101. 회귀 분석(Regression Analysis): 종속 변수와 하나 이상의 독립 변수 간의 관계를 모델링하는 데 사용되는 통계적 방법입니다.
102. 정규화(Regularization): 머신러닝 모델의 과적합을 방지하기 위해 비용 함수에 페널티 항을 추가하는 것입니다.
103. 재샘플링(Resampling): 모델의 성능을 평가하기 위한 부트스트래핑 또는 교차 검증과 같은 기술입니다.
104. ROC 곡선(ROC Curve - Receiver Operating Characteristic Curve): 분류 모델에서 다양한 임계값에 대한 참 양성률과 거짓 양성률 간의 트레이드오프를 시각적으로 표현한 것입니다.
105. 평균 제곱근 오차(RMSE - Root Mean Square Error): 예측값과 실제 값 간의 차이를 측정하는 지표입니다.
106. R-제곱(R-squared): 회귀 모델에서 종속 변수의 분산 중 독립 변수에 의해 설명되는 비율을 나타내는 통계적 측정값입니다.
107. 샘플링 편향(Sampling Bias): 결과의 일반화 가능성에 영향을 미칠 수 있는 참가자 또는 데이터 포인트 선택의 편향입니다.
108. 샘플링(Sampling): 더 큰 데이터셋에서 데이터 포인트의 하위 집합을 선택하는 과정입니다.
109. 확장성(Scalability): 증가하는 양의 데이터 또는 워크로드를 처리할 수 있는 시스템의 능력입니다.
110. 시그모이드 함수(Sigmoid Function): 이진 분류 문제에 사용되는 수학 함수입니다.
111. 실루엣 점수(Silhouette Score): 클러스터링 기법의 품질을 계산하는 데 사용되는 지표입니다.
112. 특이값 분해(SVD - Singular Value Decomposition): 차원 축소에 사용되는 행렬 분해 기술입니다.
113. 스피어만 순위 상관(Spearman Rank Correlation): 두 변수 간의 비모수적 상관 관계 측정값입니다.
114. 표준 편차(Standard Deviation): 값 집합의 변동 또는 분산 정도를 측정하는 지표입니다.
115. 정상성(Stationarity): 통계적 속성이 시간 경과에 따라 일정하게 유지되는 시계열 데이터의 속성입니다.
116. 계층 샘플링(Stratified Sampling): 모집단 내 하위 그룹의 비례적 대표성을 보장하는 샘플링 방법입니다.
117. 지도 학습(Supervised Learning): 알고리즘이 입력-출력 쌍으로 구성된 데이터셋으로 훈련되는 레이블이 지정된 데이터로부터 학습하는 것입니다.
118. 서포트 벡터 머신(SVM - Support Vector Machine): 분류 및 회귀 분석에 사용되는 지도 머신러닝 알고리즘입니다.
119. t-분포(t-Distribution): 표본 크기가 작거나 모집단 표준 편차를 알 수 없을 때 가설 검정에서 사용되는 확률 분포입니다.
120. 시계열 분석(Time Series Analysis): 시간 경과에 따라 수집된 데이터를 분석하여 패턴과 추세를 식별하는 것입니다.
121. t-검정(t-test): 두 그룹의 평균 간에 유의미한 차이가 있는지 확인하는 데 사용되는 통계 테스트입니다.
122. 두 표본 t-검정(Two-sample t-test): 두 개의 독립적인 표본의 평균을 비교하는 데 사용되는 통계 테스트입니다.
123. 과소적합(Underfitting): 데이터의 기본 패턴을 포착하기에는 너무 단순한 모델입니다.

124. 단변량 분석(Univariate Analysis): 데이터셋에서 단일 변수의 변동을 분석하는 것입니다.
125. 비지도 학습(Unsupervised Learning): 알고리즘이 자체적으로 패턴과 관계를 식별하는 레이블이 없는 데이터로부터 학습하는 것입니다.
126. 검증 세트(Validation Set): 훈련 중에 모델의 성능을 평가하는 데 사용되는 데이터의 하위 집합입니다.
127. 분산(Variance): 값 집합의 퍼짐 또는 분산 정도이며, 모델 예측의 가변성이기도 합니다.
128. XGBoost: 속도와 성능을 위해 설계된 경사 부스팅 결정 트리를 위한 오픈 소스 라이브러리입니다.
129. 제로샷 학습(Zero-shot Learning): 명시적인 예시 없이 작업을 수행하도록 모델을 훈련하는 것입니다.
130. Z-점수(Z-Score): 데이터 포인트가 평균에서 표준 편차의 몇 배만큼 떨어져 있는지를 나타내는 표준화된 점수입니다.